1. Project Title & Acronym and Abstract

Title: Microcomparative Morphosyntax Research Tool
Acronym: MIMORE
Abstract: The demonstrator tool MIMORE will be based on three databases: (i) DynaSAND (http://www.meertens.knaw.nl/sand/), a corpus of elicited speech and text collected between 2000-2005 to chart the **syntactic variation at the clausal level** in 267 dialects of Dutch spoken in the Netherlands, Belgium and North-West France; (ii) DiDDD, a corpus of elicited speech and text collected between 2005-2009 to chart the **syntactic variation at the level of nominal groups** in the same langauge area; (iii) MAND (http://www.meertens.knaw.nl/mand/database/) a corpus of elicited speech and text collected between 1980 and 1995 to chart **morphological (word-level) variation**. In the proposed tool the three databases will have a common search engine that makes it possible to investigate potential correlations between variables at the three different linguistic levels, cartographic functionality enabling the user to visualize these correlations and statistical functionality to analyze them.
Target Start Date: 02-01-2010
Target End Date: 30-06-2010
Type: Demonstrator Project

2. Coordinator

Name: prof. dr. Sjef Barbiers
Function: researcher / head of dep. variation linguistics / prof of variation linguistics
Organization: Meertens Institute (Royal Netherlands Academy of Arts and Sciences) and Utrecht University
Address: Meertens Institute; PO Box 94264, 1090 GG Amsterdam
E-mail: sjef.barbiers@meertens.knaw.nl
Tel: 020-4628530
Fax: 020-4628555
Role(s): User, Data Provider (SAND/MAND)

3. Composition of the Project Team

Name: prof. dr. Norbert Corver
Function: professor of Dutch Linguistics
Organization: Utrecht University (UiL-OTS)
Address: Trans 10, 3512 JK Utrecht
E-mail: N.F.M.Corver@uu.nl
Tel: 030-2538240
Role(s): User, Data Provider (DiDD)

Name: Kunst, drs. J.P.
Function: IT developer
Organization: Meertens Institute (Royal Netherlands Academy of Arts and Sciences)
Address: Meertens Institute; PO Box 94264, 1090 GG Amsterdam
E-mail: jan.pieter.kunst@meertens.knaw.nl

Tel: 020-4628509
Fax: 020-4628555
Role(s): Technology Provider

Name: NK
Function: technical assistant 1

Name: NK
Function: technical assistant 2

4. CLARIN centre

CLARIN Centre: Meertens Institute (Royal Netherlands Academy of Arts and Sciences). S. Barbiers and J.P. Kunst are affiliated with this centre.

5. Requested Budget

€ 55,250

6. Description of the Proposed Project

6.1 Research Question(s)

The three databases to be integrated were built up with two major goals in mind: (i) an intra-linguistic goal: to investigate theoretical questions concerning the language system and language variation; (ii) an extra-linguistic goal: to investigate the geographic distribution of (morpho-)syntactic variables.The relation between and similarity of the syntax of nominal groups and the syntax of clauses has been at the heart of formal syntactic research from the 1950's onwards. The tool will make it, for the first time, possible to systematically investigate this relation for hundreds of dialects at a microscopic level. Similarly, the relation between morphosyntactic feature specification and syntactic variation, which has been central to the so called Minimalist Program from 1992 onwards, can now be investigated systematically and microscopically. Both the intralinguistic and the extralinguistic goal will be served by a second piece of functionality, a cartographic tool that enables the user to depict one or more variables on geographic maps, thus visualizing potential correlations and associations among and between variables at three different linguistic levels. Finally, the tool should make statistical analysis of the data possible.

6.2. Research Data

	Data	Current state	Meta data	Documentation	Annotation	IPR	Participant's familiarity
Dyna-SAND	Written & spoken data from postal, oral and telephone interviews + bibliography. Transcriptions of oral interviews have been lined up with speech.	Available, searchable, mappable via web application	Name, address, sex, age, date & place of birth, education, transcriber code. Metadata in separate database, available as CIMDI; should be made accessible for search engine (except name and street).	User guide on www.meertens.nl/sand Various articles on research and database design (see references [1]. [2], [3])	POS-tagging (partial), keywords	Inform-ants have signed for permis-sion to make these data avail-able	Barbiers was project leader of DynaSAND, developed question-naires and was co-author of two atlasses based on the data set. Kunst has designed and built the database and tool.
DiDD	Written & spoken data from literature, postal and oral interviews.	Written data and trans-criptions of oral inter-views avail-able, search-able via web application; Transcript-ions not lined up with speech. Sound recordings are digitally available.	Same as DynaSAND	Reference [6]	Keywords for the data from the literature	Inform-ants have signed for permis-sion to make these data avail-able	Corver is project leader of DiDD and co-developed the question-naires. Kunst has designed and built the database and tool.
MAND	Spoken data and transcriptions (K-IPA and IPA) from oral interviews.	Transcript-ions available and search-able via web application. Sound recordings on CD-ROM	Name, address, sex, age, profession	http://www.meertens.knaw.nl/projecten/mand/. Ref. [4], [5]	POS-tagging (partial; will be converted with script)	Inform-ants have signed for permis-sion to make these data avail-able	Kunst was involved in the redesign of the MAND-database / tool

6.3 Technology

	Technology used	Current state	Meta data	Documentation	Auxiliary resources	IPR
DynaSAND	Relational database (MySQL), web application built in PHP, .mov sound files served by Darwin Streaming Server, XML-RPC web service, everything hosted on Red Hat Linux	Web application; web service in development	Not available	Minimal documentation for web service. To be developed.	None	All software and components are open source
DiDD	Relational database (MySQL), web application built in PHP, XML-RPC web service, everything hosted on Red Hat Linux	Web application (only available to project members); web service in development	Not available	Minimal documenation for web service. To be developed.	None	All software and components are open source
MAND	Relational database (MySQL), web application built in PHP, everything hosted on Red Hat Linux	Web application	Not available	Not available. To be developed.	None	All software and components are open source

6.4 Description

See 1.1: Common search engine makes it possible to search the three databases in a uniform way, cartographic tool makes it possible to visualize potential correlations between (morpho-)syntactic variables and (export to) a statistical tool makes statistical analysis possible.

6.5 Plan

Type: Demonstrator Project

Demonstrator project:
The tool should have the following ingredients:

(i) A user-friendly interface which gives a brief description of the type of data that are accesible via the tool, help functions and a manual

(ii) A common search engine:
Input: strings of orthographic characters, strings of POS-tags (and features of POS-tags), positive/negative answers to questions, keywords (if necessary) and IPA (only for MAND-data). POS tags of DynaSAND and MAND must be uniformized and conform to/be translatable into IsoCAT. DiDDD transcriptions must be tagged automatically and then corrected semi-automatically.
Output: data from the interviews; text, transcriptions and sound fragments. Accessibility of sound fragments via transcriptions. Transcriptions of DiDDD and MAND must be lined up with sound recordings (if feasible). Transcriptions of DiDDD must be marked as such (to distinguish them from answers to written questionnaires)

(iii) A web service interface to the search engine

(iv) Cartographic tool: the cartographic tool in the DynaSAND application can be used for this purpose.

(v) Statistical tool: tool for simple statistics and function for user-friendly export of data to existing statistical programmes such as SPSS.

(vi) The three databases, made compatible with the CLARIN-Infrastructure.

Demonstration scenario: A series of screen shots starting with the opening page, an example of a search (e.g., find all dialects that have *–st* suffix in 2person singular, *du*/*doe*/*dou* as a 2pS subject and all nominal groups that contain 2pS pronouns, the result of this search, the map that is drawn on the basis of it and statistical calculations on the basis of this.

Core component will be an MVC-structured set of PHP scripts. A central search component which will take the incoming search request and send it off to individual search components for each database. Each database will have its own search component which knows how to translate the incoming search request to its particular structure. The individual search components will return its search results to the central search component, which will send the combined search result to the view component for presentation to the user. The mapping component will be separate from the demonstrator, the existing Meertens "Kaart" module will be called when needed.

A separate component will be developed for **a web service interface** to the demonstrator. This component will take care of translating incoming web service requests to a format understood by the central search component, and of translating the results back to the format needed by the web service. The web service will accept and return CLARIN-compliant XML.

API for the central search component will be a set of "search" functions (searchText, searchTags, searchKeywords), which will expect strings or structs containing strings as their parameters and which will return the results as structs containing strings (for textual data) or binary data (e.g. for picture files of maps). Sounds should probably be returned in the form of URLs to the sound in question.
Existing technology and functionality can either be used as is (e.g. the mapping component) or modularized and extended to be used by the demonstrator project.

	Persons involved	Effort	Lead time	Required IS expertise	Technologies used	Auxiliary resources	Tests planned
Core component	Kunst, technical assistants	1200 (see below)	30-05-2010	Help with web service needed	MySQL, PHP, XML	Audio server	Unit testing
Application (end user interface)	Kunst, technical assistants	480	30-05-2010		PHP, JavaScript	Audio server	Acceptance testing
Demonstration scenario	Kunst, technical assistants	40	30-06-2010				
Documentation	Kunst, technical assistants	120	30-06-2010				
Making available	Kunst, technical assistants	40	30-06-2010	Help with correctly making available needed			Testing if correctly made available for CLARIN infrastructure
ISOcat mapping	Kunst, technical assistants	40	28-02-2010				

Core component
Adding PoS tags to MAND 160
Adding PoS tags to DiDDD (help from CGN-tagger needed) 160
Modularizing SAND search component 120
Modularizing MAND search component 120
Modularizing DiDDD search component 120
Creating central search component 240
Creating web service interface 280

7. Deliverables and Milestones

1. Document describing requirements and desiderata for the CLARIN infrastructure, justified by the findings of the current project (document). Barbiers, Corver & Kunst. 30-05-2010
2. Metadata of the resources dealt with in the project (data). 01-02-2010. Technical assistants & Kunst
3. Metadata made available on a recognized CLARIN server, including PIDs for the resources (milestone). 01-03-2010 (Kunst & technical assistants).
4. The application and core component underlying the demonstrator (data or software). entire team; 30-05-2010.
5. Demonstrator made available on a recognized CLARIN centre (milestone). Kunst; 30-06-2010
6. Documentation of the demonstrator application (document). (user documentation, API documentation and developer documentation). Metadata for the tool will be made available according to CLARIN standards. (entire team). (30-06-2010)

7. ISOcat extended with new entries in the user space (milestone). technical assistants. 28-02-2010.
8. Speech files and transcriptions DiDD and MAND lined up (milestone). Kunst, technical assistants. 30-4-2010.
9. SAND and MAND data uniformly tagged. 31-03-2010. technical assistants, Kunst
10. Tags assigned to DiDD transcriptions and texts. (30-05-2010). technical assistants, Kunst, IS.
11. Mapping table defining a mapping between the resource-specific linguistic categories and ISOcat data categories (data). Technical assistants. 31-03-2010.
12. Demonstration Scenario (document). 30-06-2010. Barbiers, technical assistant

8. IPR and Ethical Issues: Risks

No such risks.

9. Expertise of the applicant(s)

Jan Pieter Kunst designed, developed and build the databases and software tools for DynaSAND and DiDDD, and partially for MAND. He is currently involved in the European Dialect Syntax project, in which his main task is to develop a common search engine for a network of dialect syntax databases.

Sjef Barbiers was project leader of DynaSAND, developed its questionnaires, tagging system, transcription protocol and was the linguist primarily involved in the development of the DynaSAND tool. He is co-author of two atlasses on the basis of DynaSAND and published a number of theoretical linguistic papers on the basis of this material. He is currently the leader of the European Dialect Syntax project.

Norbert Corver (professor of Dutch Linguistics, specialization Comparative Syntax) is the project leader of DiDDD and a specialist in the syntax of nominal groups.

10. Project budget details

Participant	**Organization**	**Effort (PM)**	**Salary Costs/PM (Euro)**	**Salary Costs (Euro)**	**Travel & subsistence (Euro)**	**Total (Euro)**
Technical assistant	Meertens Institute	6	4,000	24.000	750	**24,750**
Technical assistant	Meertens Institute	6	5,000	30.000	500	**30,500**
				54.000	1,250	**55,250**
Barbiers	MI	PM	PM	0	0	**0**
Corver	UU	PM	PM	0	0	**0**
Kunst	MI	PM	PM	0	0	**0**
Total		**12**		**54,000**	**1,250**	**55,250**

11. Literature

[1] Barbiers, S., Bennis, H., Devos, M., Vogelaer, G. de, Ham, M., van der, 2005. *Syntactic Atlas of the Dutch Dialects* (SAND) Volume 1. Amsterdam University Press, Amsterdam.

[2] Barbiers, S., J. van der Auwera, E. Boef, H. Bennis, G. De Vogelaer & M. van der Ham (2008). *Syntactische Atlas van de Nederlandse Dialecten, deel II / Syntactic Atlas of the Dutch Dialects, volume II*. Amsterdam: Amsterdam University Press.

[3] Barbiers, S. & L. Cornips & J.P. Kunst (2007) 'The Syntactic Atlas of the Dutch Dialects: A corpus of elicited speech and text as an on-line dynamic atlas.' In: J.C. Beal & K.C. Corrigan & H. Moisl [red.] *Creating and digitizing language corpora. Volume 1: Synchronic databases*. Palgrave Macmillan, Hampshire, pp. 54-90.

[4] De Schutter, G., Berg, B. van den, Goeman, T. and T. de Jong (2005). *Morphological Atlas of the Dutch Dialects*. Volume 1. Amsterdam: Amsterdam University Press.

[5] Goeman, T., M. van Oostendorp, P. van Reenen, O. Koornwinder, B. van den Berg (2008). *Morphological Atlas of the Dutch Dialects* Volume 2. Amsterdam: Amsterdam University Press.

[6] Koppen, M. van, Corver, N., Kranendonk, H. and M. Rigterink (2007). The Noun Phrase: Diversity in Dutch DP Design (DiDDD). *Nordlyd* 34. http://www.ub.uit.no/baser/nordlyd/viewissue.php?id=11